AD-A118 866 ILLINOIS UNIV AT CHICAGO CIRCLE DEPT OF MATHEMATICS F/G 12/1
APPLICATION OF THE CONDITIONAL POPULATION-MIXTURE MODEL TO IMAG--ETC(U)
AUG 82 S L SCLOVE N00014-80-C-0408
UNCLASSIFIED UICC-MATH-TR-82-5 NL

1 of 1
AD
A118866

END
DATE
FILMED
10-82
DTIC

(12)

AD A118866

APPLICATION OF THE CONDITIONAL POPULATION-MIXTURE MODEL TO IMAGE SEGMENTATION

by

STANLEY L. SCLOVE

Departments of Mathematics and Quantitative Methods
University of Illinois at Chicago Circle

TECHNICAL REPORT NO. 82-5
August 15, 1982

Revision of Technical Report No. 80-1, August 15, 1980

PREPARED FOR THE
OFFICE OF NAVAL RESEARCH
UNDER
CONTRACT N00014-80-C-0408,
TASK NR042-443
with the University of Illinois at Chicago Circle

Development of Procedures and Algorithms for
Pattern Recognition and Image Processing
based on Two-Dimensional Markov Models

Principal Investigator: Stanley L. Sclove

QUANTITATIVE METHODS DEPARTMENT
COLLEGE OF BUSINESS ADMINISTRATION
UNIVERSITY OF ILLINOIS AT CHICAGO CIRCLE
BOX 4348, CHICAGO, IL 60680

8/29/82

APPLICATION OF THE CONDITIONAL POPULATION-MIXTURE MODEL TO IMAGE SEGMENTATION

STANLEY L. SCLOVE

Departments of Mathematics and Quantitative Methods
University of Illinois at Chicago Circle

CONTENTS

APPLICATION OF THE CONDITIONAL POPULATION-MIXTURE MODEL TO IMAGE SEGMENTATION

STANLEY L. SCLOVE
Departments of Mathematics and Quantitative Methods
University of Illinois at Chicago Circle

Address:
Quantitative Methods Department
College of Business Administration
University of Illinois at Chicago Circle
Box 4348, Chicago, IL 60680

ABSTRACT

The problem of image segmentation is considered in the context of a mixture of probability distributions. The segments fall into classes. A probability distribution is associated with each class of segment. Parametric families of distributions are considered, a set of parameter values being associated with each class. With each observation is associated an unobservable label, indicating from which class the observation arose. Segmentation algorithms are obtained by applying a method of iterated maximum likelihood to the resulting likelihood function. A numerical example is given. Choice of the number of classes, using Akaike's information criterion (AIC) for model identification, is illustrated.

Key words and phrases: Image processing, image segmentation, pixel classification; pattern recognition; mixtures of distributions; cluster analysis, isodata procedure, k-means procedure; Mahalanobis distance, multivariate statistical analysis; relaxation methods

APPLICATION OF THE CONDITIONAL POPULATION-MIXTURE MODEL TO IMAGE SEGMENTATION

by

STANLEY L. SCLOVE

Departments of Mathematics and Quantitative Methods
University of Illinois at Chicago Circle

I. INTRODUCTION

A digital (i.e., numerical) image may be considered as a rectangular array of picture elements (pixels), indexed by (i,j). At each pixel the same p features are observed. We denote the features by

$$X_1, X_2, \ldots, X_p.$$

The vector of features is

$$\underset{\sim}{X} = (X_1, X_2, \ldots, X_p).$$

The observed digital image is

$$\{\underset{\sim}{x}_{ij}, \ i=1,2,\ldots,I, \ j=1,2,\ldots,J\},$$

where

$$\underset{\sim}{x}_{ij} = (x_{1ij}, x_{2ij}, \ldots, x_{pij})$$

is the vector of numerical values of the p features at pixel (i,j).

Examples. (i) In color television, p = 3 colors, the pixels are the dots on the screen, and for pixel (i,j), x_{1ij} = red level, x_{2ij} = green level, and x_{3ij} = blue level. (ii) In LANDSAT data, p=4 spectral channels, one in the green/yellow visible range, the second in the red visible range, and the other two in the near infrared range.

An object is a set of contiguous pixels which may be assumed to be members of a common class. One task of image processing is segmentation, grouping of pixels with a view toward identifying objects.

In this context the conceptual model is that the image is a set of pixels, and, also, the image consists of several segments. Each pixel belongs to one and only one segment. The segments fall into several classes. For example, in a picture of a house the classes might be brick, sky, grass, shadow and brush. Note that there might be several separate areas of, say, grass. Each of these areas is a segment, but they all belong to the class, "grass."

The statistical model accompanying this conceptual model is as follows:

--With each class of segment is associated a probability distribution for the feature vector $\underset{\sim}{X}$;

--With each pixel is associated a label which, were it known to us, would tell us which class of segment the pixel belongs to.

Each pixel thus gives rise to a pair $(\underset{\sim}{X},\gamma)$, where $\underset{\sim}{X}$ is observable and γ is not. In the context of this statistical model segmentation is estimation of the set of labels.

The number of classes will be denoted by k. The algorithms developed here try one value of k at a time. Methods of comparing the results for different values of k will be discussed.

Often one considers parametric models, in which the class-conditional probability functions $f(\underset{\sim}{x}|c)$ are assumed known, except possibly for the values of distributional parameters. That is,

$$f(\underset{\sim}{x}|c) = h(\underset{\sim}{x};\underset{\sim}{\beta}_c),$$

where $\underset{\sim}{\beta}_c$ is the parameter. E.g., in the multivariate Gaussian case $\underset{\sim}{\beta}_c$ consists of the mean and covariance matrix for class c. The parameters are usually unknown. However, image processing is usually done in a context where there is prior information about the parameters. This can provide initial estimates for an iterative estimation algorithm.

We shall write $\underset{\sim}{x}_t$ rather than $\underset{\sim}{x}_{ij}$, using a single subscript t rather than the double subscript ij for the pixels, even though they are in a two-dimensional array.

The label associated with the t-th pixel will be denoted by γ_t, $t = 1,2,\ldots, n = IJ$. The label is equal to c if and only if pixel t belongs to class c. It is convenient to represent the information carried by the label in a k-dimensional vector $\underset{\sim}{\theta}_t$ which consists of k-1 zeros and a single 1, the position of the 1 indicating which segment pixel t belongs to; i.e., $\underset{\sim}{\theta}_t$ has a 1 as its γ_t-th element and 0's elsewhere. The probability density function (p.d.f.) of $\underset{\sim}{X}_t$, given $\underset{\sim}{\theta}_t$, is

$$f(\underset{\sim}{x}_t|\underset{\sim}{\theta}_t) = \Sigma_c \theta_{ct} f(\underset{\sim}{x}_t|c), \qquad (1.1)$$

where the summation is for $c = 1,2,\ldots,k$, and θ_{ct} is the c-th element of $\underset{\sim}{\theta}_t$.

II. THE PROBABILITY MODEL

It is assumed that the $\underset{\sim}{X}$'s are conditionally independent, given the γ's. (More complicated models are under study.) Then their joint p.d.f. is the product over $t = 1,2,\ldots,n$ of factors (1.1).

Note that, if $\underset{\sim}{X}_1, \underset{\sim}{X}_2, \ldots, \underset{\sim}{X}_n$ are independent and identically distributed with a standard mixture density

$$f(\underset{\sim}{x}) = \Sigma_c f(\underset{\sim}{x}|c)\pi_c,$$

where the summation is for $c = 1,2,\ldots,k$ and the sum of the class probabilities π_c is 1, then (1.1) gives the conditional density of the $\underset{\sim}{X}$'s, given their labels. It is for this reason that the model used here is called the <u>conditional</u> population-mixture model. The standard mixture model has been used for pixel classification; see, e.g., [7]. Further discussion of the conditional model, in the context of statistical cluster analysis, and further references are given in [10].

A likelihood approach is illuminating in that it can show how *ad hoc* optimality criteria (objective functions) which have been proposed relate to likelihood function in particular probability models.

Note that (1.1) can be written as a product

$$f(\underset{\sim}{x}_t | \underset{\sim}{\theta}_t) = \Pi_c \, f(\underset{\sim}{x}_t | c)^{\theta_{ct}}, \qquad (2.1)$$

where the product is over c = 1,2,...,k. This form is often more convenient, and we shall use it in what follows.

III. THE SEGMENTATION ALGORITHM

Using (2.1) and the conditional independence assumption, one sees that the joint p.d.f. of the $\underset{\sim}{X}_t$, given the $\underset{\sim}{\theta}_t$, is

$$\Pi_t \Pi_c \, [h(\underset{\sim}{x}_t ; \underset{\sim}{\beta}_c)]^{\theta_{ct}} .$$

This likelihood is to be maximized over all assignments of pixels to classes and over all permissible parameter values. Many *ad hoc* schemes can be applied to this maximization problem. E.g., one way to maximize is to start with a given segmentation, take each observation successively and shift it to the first segment for which a shift results in an increase in likelihood, and loop through the data until no pixel changes classes.

The algorithm developed here is an iterative, back-and-forth procedure. We first maximize with respect to (w.r.t.) the $\underset{\sim}{\theta}$'s (holding the $\underset{\sim}{\beta}$'s fixed at initial values), then w.r.t. the $\underset{\sim}{\beta}$'s (holding the $\underset{\sim}{\theta}$'s fixed at the values obtained in the previous stage), then again w.r.t. the $\underset{\sim}{\theta}$'s (holding the $\underset{\sim}{\beta}$'s fixed at the values obtained in the previous stage), etc. We stop when no $\underset{\sim}{\theta}$ changes, i.e., when no pixel changes classes, or when a specified amount of computer time is used or a specified number of iterations has been performed.

An alternative way of starting the procedure would be to start with an initial segmentation rather than with initial guesses of the β's.

It is clear that, for fixed values of the β's, say b's, the likelihood is maximized, for each t, by taking the estimate T_{ct} of θ_{ct} to be

$$T_{ct} = \begin{cases} 1 & \text{if } h(x_t; b_c) = \max_d\{h(x_t; b_d)\} \\ 0 & \text{otherwise .} \end{cases} \tag{3.1}$$

(In case of ties an arbitrary choice is made; e.g., the observation is assigned to the tieing class with smallest subscript.) In other words, segmentation proceeds by allocating pixel t to that class c for which the estimated probability density of the observation x_t is largest.

Note that, having tentatively estimated the θ's at any stage, i.e., having tentatively segmented the image, estimation of the β's is reduced simply to ordinary maximum likelihood estimation in the particular parametric family at hand. *This is a special advantage of this approach.*

Let Θ denote the set of θ's and B the set of β's. Let $L(B, \Theta)$, or simply L for short, denote the likelihood. Let $B^{(s)}$ denote the value of B which maximizes L at the s-th stage of the iteration, and let $\theta^{(s)}$ denote the value of Θ which maximizes L at the s-th stage of the iteration. Then $\theta^{(s)}$ maximizes $L(B^{(s)}, \Theta)$ w.r.t. Θ, and $B^{(s)}$ maximizes $L(B, \theta^{(s-1)})$ w.r.t. B. This back-and-forth maximization is an example of the <u>relaxation method</u> (<u>Southwell's method</u>); see [7, pp. 241ff.] and [10,11]. It is true that

$$L(B^{(s+1)}, \theta^{(s)}) \geq L(B^{(s)}, \theta^{(s)})$$

and

$$L(\underset{\sim}{B}^{(s)}, \underset{\sim}{\theta}^{(s+1)}) \geq L(\underset{\sim}{B}^{(s)}, \underset{\sim}{\theta}^{(s)}) .$$

That is, at no stage of the procedure can the value of the likelihood decrease; however, there is no guarantee of convergence to the global maximum (neither do alternative clustering algorithms guarantee convergence to the global maximum of their objective functions). To see how the procedure can fail to converge to a global maximum, suppose it happens that

$$L(\underset{\sim}{B}^{(s)}, \underset{\sim}{\theta}^{(s)}) > L(\underset{\sim}{B},\underset{\sim}{\theta}^{(s)}) \quad \text{for all } \underset{\sim}{B},$$

or

$$L(\underset{\sim}{B}^{(s)}, \underset{\sim}{\theta}^{(s-1)}) > L(\underset{\sim}{B}^{(s)}, \underset{\sim}{\theta}) \quad \text{for all } \underset{\sim}{\theta}..$$

Then the procedure will terminate at the s-th stage, without having necessarily reached the global maximum. That is, if, having maximized w.r.t. one of the variables $\underset{\sim}{B}$ or $\underset{\sim}{\theta}$, we happen to find ourselves at a (relative) maximum w.r.t. the other, we may not reach a global maximum. In other words, the procedure could conceivably stop at a multidimensional saddle point.

IV. APPLICATION TO PARTICULAR DISTRIBUTIONS

Now we consider application of this general method to particular families of distributions. First we consider normal distributions with common covariance matrix, for in this case it becomes clear how the model of the present paper establishes a link with some existing clustering procedures.

A. Multivariate Normal Distributions with Common Covariance Matrix

In the case of normal distributions with means $\underset{\sim}{\mu}_c$, $c = 1,2,\ldots,k$, and common covariance matrix $\underset{\sim}{\Sigma}$, the likelihood takes the form

$$(2\pi)^{-np/2}|\underset{\sim}{\Sigma}|^{-n/2}\exp[-\Sigma_t\Sigma_c\theta_{ct}q(\underset{\sim}{x}_t;\underset{\sim}{\mu}_c,\underset{\sim}{\Sigma})/2],$$

where the quadratic form q is given by

$$q(\underset{\smile}{x};\underset{\smile}{\mu},\underset{\smile}{\Sigma}) = (\underset{\smile}{x} - \underset{\smile}{\mu})'\underset{\smile}{\Sigma}^{-1}(\underset{\smile}{x} - \underset{\smile}{\mu}),$$

where ' denotes vector transpose. This quadratic form is the squared (Mahalanobis) distance between $\underset{\smile}{x}$ and $\underset{\smile}{\mu}$ in the metric of $\underset{\smile}{\Sigma}$. Here (3.1) is equivalent to

$$T_{ct} = \begin{cases} 1 & \text{if } q(\underset{\smile}{x}_t;\underset{\smile}{m}_c,\underset{\smile}{S}) = \min_d\{q(\underset{\smile}{x}_t;\underset{\smile}{m}_d,\underset{\smile}{S})\} \\ 0 & \text{otherwise,} \end{cases} \tag{4.1}$$

where $\underset{\smile}{m}_c$ and $\underset{\smile}{S}$ are, respectively, the estimates of $\underset{\smile}{\mu}_c$ and $\underset{\smile}{\Sigma}$. That is, pixel t is assigned to that group to whose tentatively estimated mean vector it is closest, where the distance is in the metric of the tentatively estimated covariance matrix. Having estimated the $\underset{\smile}{\theta}$'s, we have multivariate normal observations arranged into groups; maximization w.r.t. the $\underset{\smile}{\mu}$'s and $\underset{\smile}{\Sigma}$ is accomplished by taking the group mean vectors as estimates of the $\underset{\smile}{\mu}$'s, and the within-groups sum-of-products matrix gives the estimate of $\underset{\smile}{\Sigma}$. The procedure is iterated: using new estimates $\underset{\smile}{m}_c$, $c = 1,2,\ldots,k$, and $\underset{\smile}{S}$, the rule (4.1) is applied again. Then new $\underset{\smile}{m}$'s and a new $\underset{\smile}{S}$ are calculated; etc. The Mahalanobis distances can be computed efficiently; see, e.g., [1, p. 107].

Relation to the isodata procedure: This scheme is a Mahalanobis distance version of isodata [4]. Isodata proceeds as follows. One starts with tentative estimates of cluster means and assigns each individual to the mean to which he is closest. (Isodata uses Euclidean distance, or modified Euclidean distance in which different weights are assigned to the p dimensions.) The cluster means are then re-estimated, and one loops through the data again, reassigning the individuals, etc. Note the similarity to our scheme: We start with tentative estimates of the means and covaraiance matrix and assign each individual to the mean to which he is closest, using Mahalanobis distance in the metric of the tentatively estimated covariance matrix. The means and covariance matrix are then re-estimated, the individuals (pixels) are re-allocated to clusters (segment classes), etc.

An important difference is that our scheme employs Mahalanobis distance rather than Euclidean or weighted-Euclidean distance. It is worth emphasizing that it is the Mahalanobis distance based on the within-groups sum-of-products matrix that arises here; some data analysts use the total sum-of-products matrix, which is not appropriate; see, e.g., [5]. Thus, if one wants to achieve use of a proper metric by making a linear transformation of the data, this would have to be done at the beginning of each iteration, making the appropriate transformation based on the covariance matrix estimate obtained at the previous iteration.

Relation to the k-means procedure: Arranging the computation differently, updating the estimates of the means and covariance matrix after each individual pixel is assigned rather than waiting until all have been assigned, produces a Mahalanobis-distance version of the k-means procedure [8].

A numerical example: As a sample "image" the Fisher iris data were used. This dataset consists of 4 features measured on 150 flowers, 50 in each of three species. To form a digital image the 150 flowers were arranged into a 15 x 10 rectangular array, rows 1-5 being species 1, rows 6-10 being species 2, rows 11-15 being species 3. This means that the true segmentation is as follows. (Note that, although these data are arranged in a rectangular array, no use was made of the spatial information. Paper [11] is a preliminary report of the development of algorithms incorporating spatial and contextual information.)

TRUE SEGMENTATION:

ROW:	COLUMN: 1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1	1	1
6	2	2	2	2	2	2	2	2	2	2
7	2	2	2	2	2	2	2	2	2	2
8	2	2	2	2	2	2	2	2	2	2
9	2	2	2	2	2	2	2	2	2	2
10	2	2	2	2	2	2	2	2	2	2
11	3	3	3	3	3	3	3	3	3	3
12	3	3	3	3	3	3	3	3	3	3
13	3	3	3	3	3	3	3	3	3	3
14	3	3	3	3	3	3	3	3	3	3
15	3	3	3	3	3	3	3	3	3	3

Below are given results obtained by starting with initial means equal to the measurements on flowers 50, 100 and 150. (These are easy for the algorithm in the sense that they are in fact from the three different species, but not so easy as, e.g., flowers 1, 51 and 101, which are further apart. Starting with means that are from correct classes is analogous to applications where something is known about the characteristics of the classes.) The results in successive iterations were as follows. Convergence was reached on the fourth iteration, i.e., on the fifth iteration no pixel changed class. The execution time was 8.81 sec. on an IBM 4341.

SEGMENTATION ON ITERATION 1:

ROW:	COLUMN: 1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1	1	1
6	3	3	3	2	3	2	3	2	3	2
7	2	2	2	3	2	2	2	2	2	2
8	3	2	3	2	2	2	3	3	2	2
9	2	2	2	3	2	3	3	2	2	2
10	2	3	2	2	2	2	2	2	2	2
11	3	3	3	3	3	3	2	3	3	3
12	3	3	3	3	3	3	3	3	3	3
13	3	3	3	3	3	3	3	3	3	3
14	3	3	3	3	3	3	3	3	3	3
15	3	3	3	3	3	3	3	3	3	3

CONFUSION MATRIX:

		True Class 1	2	3	
Label	1	50	0	0	50
	2	0	35	1	36
	3	0	15	49	64
		50	50	50	150

16 errors

$-2 \log_e L = 258.7$

SEGMENTATION ON ITERATION 2:

ROW:	COLUMN: 1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1	1	1
6	2	2	3	2	2	2	3	2	2	2
7	2	2	2	2	2	2	2	2	2	2
8	3	2	3	2	2	2	2	3	2	2
9	2	2	2	3	2	2	2	2	2	2
10	2	2	2	2	2	2	2	2	2	2
11	3	3	3	3	3	3	2	3	3	3
12	3	3	3	3	3	3	3	3	3	3
13	3	3	3	3	3	3	3	3	3	3
14	3	3	3	3	3	3	3	3	3	3
15	3	3	3	3	3	3	3	3	3	3

CONFUSION MATRIX:

		True Class 1	2	3	
Label	1	50	0	0	50
	2	0	44	1	45
	3	0	6	49	55
		50	50	50	150

7 errors

$-2 \log_e L = 212.2$

SEGMENTATION ON ITERATION 3:

ROW:	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1	1	1
6	2	2	2	2	2	2	2	2	2	2
7	2	2	2	2	2	2	2	2	2	2
8	3	2	2	2	2	2	2	3	2	2
9	2	2	2	3	2	2	2	2	2	2
10	2	2	2	2	2	2	2	2	2	2
11	3	3	3	3	3	3	3	3	3	3
12	3	3	3	3	3	3	3	3	3	3
13	3	3	3	3	3	3	3	3	3	3
14	3	3	3	3	3	3	3	3	3	3
15	3	3	3	3	3	3	3	3	3	3

(COLUMN: 1–10)

CONFUSION MATRIX:

	True Class 1	2	3	
Label 1	50	0	0	50
Label 2	0	47	0	47
Label 3	0	3	50	53
	50	50	50	150

3 errors

$-2 \log_e L = 190.4$

SEGMENTATION ON ITERATION 4:

ROW:	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1	1	1
6	2	2	2	2	2	2	2	2	2	2
7	2	2	2	2	2	2	2	2	2	2
8	3	2	2	2	2	2	2	2	2	2
9	2	2	2	3	2	2	2	2	2	2
10	2	2	2	2	2	2	2	2	2	2
11	3	3	3	3	3	3	3	3	3	3
12	3	3	3	3	3	3	3	3	3	3
13	3	3	3	3	3	3	3	3	3	3
14	3	3	3	2	3	3	3	3	3	3
15	3	3	3	3	3	3	3	3	3	3

(COLUMN: 1–10)

CONFUSION MATRIX:

	True Class 1	2	3	
Label 1	50	0	0	50
Label 2	0	48	1	49
Label 3	0	2	49	51
	50	50	50	150

3 errors

$-2 \log_e L = 187.6$

All computations reported here were carried out using FORTRAN computer programs written by the author. These programs have been sent to the Statistics Program at the Office of Naval Research for deposit in the Naval Research Laboratories.

B. Multivariate Normal Distributions with Different Covariance Matrices

The algorithm generated for this case turns out not to be simply to use a different Mahalanobis distance for each cluster. (The complication which occurs is analogous to that in classification, i.e., discriminant analysis, where one is led to quadratic discriminant functions if the covariance matrices differ.) The likelihood is

$$(2\pi)^{-np/2} \Pi_c \Pi_t |\underset{\sim}{\Sigma}_c|^{-\theta_{ct}/2} \exp[-\Sigma_c \Sigma_t \theta_{ct} q(\underset{\sim}{x}_t; \underset{\sim}{\mu}_c, \underset{\sim}{\Sigma}_c)/2].$$

Equation (3.1) becomes

$$T_{ct} = \begin{cases} 1 & \text{if setting } d=c \text{ maximizes } |\underset{\sim}{\Sigma}_d|^{-1/2} \exp[-q(\underset{\sim}{x}_t; \underset{\sim}{\mu}_d, \underset{\sim}{\Sigma}_d)/2] \\ 0 & \text{otherwise .} \end{cases} \qquad (4.2)$$

Maximizing the expression in (4.2) is equivalent to minimizing

$$\log_e |\underset{\sim}{\Sigma}_d| + q(\underset{\sim}{x}_t; \underset{\sim}{\mu}_d, \underset{\sim}{\Sigma}_d)$$

This involves not only the Mahalanobis distance between the observation and the mean of the given class but also the logarithm of the determinant of the covariance matrix for the given class.

It has been noted (see, e.g., [6]) that in the standard mixture model for this case the supremum of the likelihood is infinity. This is reflected in the fact that in our algorithm it would be possible that at some stage one of the classes would consist of a single pixel, so that the tentative estimate of the mean of that group would be the feature vector for that pixel, and the tentative estimate of the covariance matrix of that cluster would be undefined.

Numerical example, continued: Results similar to those for the case of common covariance matrix were obtained using the algorithm for this case, with the adjustment for determinants of the covariance matrices. However, when these adjustments were omitted, and the clustering was performed using only the Mahalanobis distances, without adding the logarithm of the determinant of the covariance matrix, the results were poor. Fifty flowers were correctly assigned to class 1, but only 6 were assigned to class 2, the remaining 94 being assigned to class 3. Also, it took twelve iterations and 27 sec.'s of CPU time to obtain this poor result.

V. COMPARISON WITH THE METHOD BASED ON THE STANDARD MIXTURE MODEL

Clustering based on the standard mixture model was considered in [14]. Under that model the posterior probability that Individual t belongs to Class c is

$$\pi_c \, h(\underset{\sim}{x}_t;\underset{\sim}{\beta}_c)/\Sigma_d \, \pi_d \, h(\underset{\sim}{x}_t;\underset{\sim}{\beta}_d) \, . \qquad (5.1)$$

Individual t is assigned to that class c for which the estimate of (5.1) is largest, i.e., to that class for which the estimated posterior probability of membership is largest. On the other hand, with the conditional mixture model, Individual t is assigned to that class c for which the estimate of the density $h(\underset{\sim}{x}_t;\underset{\sim}{\beta}_c)$ is largest.

Wolfe [14] has provided computer programs for the standard mixture model in the case of normal distributions. As is well known, the likelihood equations for mixture problems are relatively complicated. In [14] they are solved by a multivariate Newton-Raphson iterative method. This involves the assignment of arbitrary initial values to start the iterative solution, as does the method described here.

VI. SOME REMARKS ON STATISTICAL INFERENCE

The maximum likelihood estimate of $(\underset{\smile}{B},\underset{\smile}{\Theta})$ is the value $(\underset{\smile}{b},\underset{\smile}{t})$ for which the likelihood L is largest. The quantity $L(\underset{\smile}{b},\underset{\smile}{t})$ is the corresponding maximum value of the likelihood. To approximate $(\underset{\smile}{b},\underset{\smile}{t})$ one uses the algorithm. Let $\Lambda(\underset{\smile}{B},\underset{\smile}{\Theta}) = L(\underset{\smile}{B},\underset{\smile}{\Theta})/L(\underset{\smile}{b},\underset{\smile}{t})$. Let F denote the large sample cumulative distribution function (c.d.f.) of $-2 \log_e \Lambda$, i.e.,

$$\lim_{n\to\infty} \Pr[-2 \log_e \Lambda(\underset{\smile}{B},\underset{\smile}{\Theta}) \le x] = F(x).$$

Suppose that F is independent of the true values $(\underset{\smile}{B},\underset{\smile}{\Theta})$. E.g., it may be the c.d.f. of a chi-square distribution with an appropriate number of degrees of freedom; it is necessary to investigate the extent to which the large sample theory of the generalized likelihood ratio applies when there are incidental parameters (such as the labels).

A. Confidence Sets

Let x_α denote the upper α-th percentage point of F. Then

$$\begin{aligned} 1-\alpha &= F(x_\alpha) = \Pr[-2 \log_e \Lambda(\underset{\smile}{B},\underset{\smile}{\Theta}) \le x_\alpha] \\ &= \Pr[-2 \log_e L(\underset{\smile}{B},\underset{\smile}{\Theta}) \le x_\alpha + 2 \log_e L(\underset{\smile}{b},\underset{\smile}{t})] \end{aligned}$$

so that

$$\{(\underset{\smile}{B},\underset{\smile}{\Theta}):\ -2 \log_e L(\underset{\smile}{B},\underset{\smile}{\Theta}) \le x_\alpha + 2 \log_e L(\underset{\smile}{b},\underset{\smile}{t})\}$$

is an approximate $100(1-\alpha)\%$ confidence set for $(\underset{\smile}{B},\underset{\smile}{\Theta})$. Denote by $(\underset{\smile}{b}',\underset{\smile}{t}')$ the estimates produced by the algorithm. Then, since there is a possibility these have not quite converged to the maximum likelihood estimates $(\underset{\smile}{b},\underset{\smile}{t})$, we have $L(\underset{\smile}{b}',\underset{\smile}{t}') \le L(\underset{\smile}{b},\underset{\smile}{t})$. Thus a conservative confidence set (one that contains more values of $(\underset{\smile}{B},\underset{\smile}{\Theta})$ than the true confidence set and thus has confidence coefficient at least $1-\alpha$) is

$$\{(\underset{\smile}{B},\underset{\smile}{\Theta}): \ -2\log_e L(\underset{\smile}{B},\underset{\smile}{\Theta}) \ \le \ x_\alpha + 2\log_e L(\underset{\smile}{b}',\underset{\smile}{t}')\}.$$

B. Some Remarks on Choice of Number of Classes

One *ad hoc* approach to the choice of number of classes is to follow the suggestion in [8] of introducing refinement and coarsening parameters R and C such that two clusters join when their mean vectors are less than C units apart and a cluster splits when its diameter exceeds R.

Another approach is to run the algorithm with different choices of k and compare the results. Note that the likelihood function is a different function for different values of k. Denote this dependence upon k by writing the likelihood as $L_k(\underset{\smile}{B}(k),\underset{\smile}{\Theta}(k))$. Let $\underset{\smile}{b}(k)$, $\underset{\smile}{t}(k)$ denote the maximum likelihood estimates for fixed k. Following the approach of [14] for the standard mixture model, one might make a sequence of hypothesis tests to decide on k, first comparing $L_2(\underset{\smile}{b}(2),\underset{\smile}{t}(2))$ with $L_3(\underset{\smile}{b}(3),\underset{\smile}{t}(3))$, then if necessary comparing $L_3(\underset{\smile}{b}(3),\underset{\smile}{t}(3))$ with $L_4(\underset{\smile}{b}(4),\underset{\smile}{t}(4))$, etc. In [14] the asymptotic chi-square distribution of the generalized likelihood ratio is used; even in the context of the standard mixture model this may not be the correct asymptotic distribution.

Still another approach is to use Akaike's information criterion (AIC). (See, e.g., [1,2].) This statistic is

$$AIC(k) \ = \ -2\log_e[\max L_k] \ + \ 2\,m(k).$$

Here $m(k)$ is the number of independent parameters estimated when using k classes. According to this viewpoint on model selection, the best model is the one that minimizes AIC. According to AIC, inclusion of an additional parameter isappropriate if $\log_e$[max L] increases by one unit or more, i.e., if max L increases by a factor of e or more.

Numerical example, continued: There is some question as to whether the Fisher iris data should be treated as two or as three species and whether separate covariance matrices should be used for the species. (See [3, pp. 109-110].) Accordingly, we compare by AIC the four models resulting from taking k=2 and 3 and using common and separate covariance matrices. The results were as follows.

Values of AIC

k	common covariance matrix	separate covariance matrices
2	437.9	293.8
3	231.6	123.3

For both values of k, the model with separate covariance matrices fared better, and k=3 gave a smaller value of AIC than did k=2.

VII. DISCUSSION

A. Conclusions

A probability framework for clustering/segmentation problems has been discussed. A general method of producing algorithms which correspond to a method of iterated maximum likelihood has been given. The general method given here is plausible, is linked to a probability model, and is easy to program. In the case of multivariate normal distributions with common covariance matrix the general method produces schemes which can be viewed as improved versions of some existing schemes.

B. Remarks

The focus here has been on the parametric case, but the methods discussed might be applied nonparametrically, by estimating the p.d.f.'s $f(\underset{\sim}{x}|c)$ as segmentation proceeds, using standard methods of density estimation.

Algorithms based on a likelihood function are based on the raw data matrix, in contrast to may clustering procedures which are based on a matrix of pairwise similarities or distances. The latter procedures have the advantage of applicability to problems where a raw data matrix is not available. When the raw data are available, such algorithms have the theoretical disadvantage of not extracting all the information from the observations and the computational disadvantage of preliminary computation of all the pairwise distances.

C. Alternative Models

The focus here has been on a model in which the labels are treated as functionally independent. In the standard mixture model they become random variables and are treated as statistically independent.
To the assumption of Section I is seems reasonable to add:

--Each segment consists of more than one pixel.

As a corollary to this assumption, it follows that the labels are functionally related, in as much as each label must be equal to one of its eight neighbors. It would be interesting to study the problem resulting from maximizing the likelihood function under this condition.
Alternatively, if the labels are then treated as random, they would be a two-dimensional Markov process. The author has developed an algorithm for

estimation in this Markov model. Paper [11] is a preliminary report on this; a more detailed report is forthcoming.

REFERENCES

[1] H. Akaike, "A new look at statistical model identification," IEEE Trans. Auto. Control, vol. AC-19, pp. 716-723, 1974.

[2] H. Akaike, "Likelihood of a model and information criteria," Journal of Econometrics, vol. 16, pp. 1-14, 1981.

[3] T. W. Anderson, An Introduction to Multivariate Statistical Analysis. New York: Wiley, 1958.

[4] G. H. Ball and D. J. Hall, "A clustering technique for summarizing multivariate data," Behavioral Science, vol. 12, pp. 153-155, 1967.

[5] H. Chernoff, "Metric considerations in cluster analysis," Proc. 6th Berkeley Symp. Math. Statist. Prob., vol. 1. Los Angeles and Berkeley: Univ. of Calif. Press, pp. 621-629, 1970.

[6] N. E. Day, "Estimating the components of a mixture of normal distributions," Biometrika, vol. 56, pp. 463-475, 1969.

[7] J. O. Eklundh, H. Yamamoto and A. Rosenfeld, "A relaxation method for multispectral pixel classification," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. PAMI-2, pp. 72-75, 1980.

[8] J. MacQueen, "Some methods for classification and analysis of multivariate observations," Proc. 5th Berkeley Symp. Math. Statist. Prob., vol. 1. Los Angeles and Berkeley: Univ. of Calif. Press, pp. 281-297, 1966.

[9] J. Ortega and W. Rheinboldt, Iterative Solution of Nonlinear Equations in Several Variables. New York: Academic Press, 1970.

[10] S. L. Sclove, "Population mixture models and clustering algorithms," Communications in Statistics (A), vol. A6, pp. 417-434, 1977.

[11] S. L. Sclove, "Segmentation of time series and images in the signal detection and remote sensing contexts," Technical Report No. 82-4, ONR Contract N00014-80-C-0408, Task NR042-443, University of Illinois at Chicago Circle. To appear in Proceedings of the Workshop on Signal Processing in the Ocean Environment, U.S. Naval Academy, Annapolis, MD, May 11-14, 1982. New York: Marcel-Dekker, Inc.

[12] R. Southwell, Relaxation Methods in Engineering Science: a Treatise on Approximate Computation. London: Oxford Univ. Press, 1940.

[13] R. Southwell, Relaxation Methods in Theoretical Physics. London and New York: Oxford Univ. Press (Clarendon), 1946.

[14] J. H. Wolfe, "Pattern clustering by multivariate mixture analysis," Multivariate Behavioral Research, vol. 5, pp. 329-350, 1970.

TECHNICAL REPORTS

OFFICE OF NAVAL RESEARCH CONTRACT N00014-80-C-0408, TASK NR042-443

with the University of Illinois at Chicago Circle

Development of Procedures and Algorithms for
Pattern Recognition and Image Processing
based on Two-Dimensional Markov Models

Principal Investigator: Stanley L. Sclove

No. 80-1. Stanley L. Sclove. "Application of the Conditional Population-Mixture Model to Image Segmentation." 8/15/80

No. 80-2. Stanley L. Sclove. "Modeling the Distribution of Fingerprint Characteristics." 9/19/80

No. 81-1. Stanley L. Sclove. "On Segmentation of Time Series." 11/30/81

No. 82-1. Hamparsum Bozdogan and Stanley L. Sclove. "Multi-Sample Cluster Analysis using Akaike's Information Criterion." 1/30/82

No. 82-2. Hamparsum Bozdogan and Stanley L. Sclove. "Multi-Sample Cluster Analysis with Varying Parameters using Akaike's Information Criterion." 3/8/82

No. 82-3. Stanley L. Sclove. "Some Aspects of Inference for Multivariate Infinitely Divisible Distributions." 6/15/82

No. 82-4. Stanley L. Sclove. "On Segmentation of Time Series and Images in the Signal Detection and Remote Sensing Contexts." 8/1/82

No. 82-5. Stanley L. Sclove. "Application of the Conditional Population-Mixture Model to Image Segmentation." 8/15/82
Revision of Technical Report No. 80-1.

8/18/82

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Technical Report 82-5	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Application of the Conditional Population-Mixture Model to Image Segmentation		5. TYPE OF REPORT & PERIOD COVERED Technical Report
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Stanley L. Sclove		8. CONTRACT OR GRANT NUMBER(s) N00014-80-C-0408
9. PERFORMING ORGANIZATION NAME AND ADDRESS University of Illinois at Chicago Circle Box 4348, Chicago, IL 60680		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS		12. REPORT DATE August 15, 1982
		13. NUMBER OF PAGES 19 + ii
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Office of Naval Research Statistics and Probability Branch Arlington, VA 22217		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Unlimited distribution

18. SUPPLEMENTARY NOTES

Revision of Technical Report 80-1

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Image processing, image segmentation, pixel classification; pattern recognition; mixtures of distributions; cluster analysis, isodata procedure, k-means procedure; Mahalanobis distance, multivariate statistical analysis; relaxation methods

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The problem of image segmentation is considered in the context of a mixture of probability distributions. The segments fall into classes. A probability distribution is associated with each class of segment. Parametric families of distributions are considered, a set of parameter values being associated with each class. With each observation is associated an unobservable label, indicating from which class the observation arose. Segmentation algorithms are obtained by applying a method of iterated

DD FORM 1473, 1 JAN 73 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

(Abstract, continued)

maximum likelihood to the resulting likelihood function. A numerical example is given. Choice of the number of classes, using Akaike's information criterion (AIC) for model identification, is illustrated.

F

10